Tag 5 · Gruppe 2

Gruppenname: Florian, Gruppe 2	**Product Owner:** _____	**Process Expert:** _____

Product Goal, Hypothese, Metrik und Schwelle

Ein Onlineshop plant einen KI-Assistenten für Versand- und Rückgabefragen mit Übergabe an Mitarbeitende. Hypothese: Mindestens 60 % aller Anfragen werden korrekt ohne Übergabe gelöst. Metrik: korrekt ohne Übergabe gelöste Anfragen / alle Anfragen im gleichen Testbestand. Die 60 % sind ein vorab gesetzter Übungswert. Eigene Pilotdaten fehlen. Die Recherche prüft externe Evidenz und Übertragbarkeit, nicht die tatsächliche Wirkung unseres noch ungetesteten Assistenten.

Data Driven Scrum – Aufträge der Agentengruppe (Create) und geprüfte Quellen (Observe)

Auftrag 1 · Messdefinitionen	Auftrag 2 · Unabhängige Erfolgsbefunde	Auftrag 3 · Gegenbelege und Fehlschläge	Auftrag 4 · Bedingungen der Wissensbasis	Auftrag 5 · Übergabe und Qualitätsprüfung
Wie wird „korrekt ohne Übergabe gelöst“ definiert und gemessen — und welche Fehlerquellen haben die Methoden?	Welche Containment-/Deflection-/Resolution-Raten sind unabhängig belegt — und wie ist die Metrik definiert?	Welche Evidenz spricht gegen hohe Automatisierung — Fehlantworten, Eskalationsversagen, Frustration?	Welche inhaltlichen und technischen Bedingungen entscheiden über die Quote?	Wie müssen Übergabe und laufende QA gestaltet und gemessen werden, damit die Quote nicht geschönt ist?
Abgrenzung: Keine Performance-Zahlen; nur Metrik- und Methodenlandkarte. Quellen geprüft: 27 Quellen / 20 Domains · USPS OIG 2021 (Containment-Definition), Brown/Cai/DasGupta 2001 (Wilson-Intervall), NIST AI 100-1	Abgrenzung: Keine Hochrechnung, keine Vendor-Cases als Evidenz. Quellen geprüft: 25 Quellen / 17 Domains · Gartner 08/2024 (n = 5.728, 14 %), Freshworks 2025 (Retail 53 %), Intercom-Fin-Pricing	Abgrenzung: Keine allgemeine KI-Kritik ohne Servicebezug. Quellen geprüft: 31 Quellen / 25 Domains · Alibaba-Feldexperiment arXiv:2605.14830 (647 Agenten, 680.676 Chats), Platforma 2026, Klarna/BI-Korrektur	Abgrenzung: Keine Modellvergleiche, keine Performance-Zahlen. Quellen geprüft: 29 Quellen / 18 Domains · DSK-OH 05/2024 (Rz. 12–13), HICSS 2021 (DP1–DP5), Gartner (43 % ohne passenden Inhalt)	Abgrenzung: Designwissen statt Effektstärken; keine Rechtsberatung. Quellen geprüft: aus Rohrecherche 6 (29 Quellen) · HICSS 2021 Übergabe-Design, DSK Rz. 12–13, Platforma 2026 (59 % erklären erneut)

Auftrag 6 · DACH-Praxis	Auftrag 7 · Intent-Verteilung	Auftrag 8 · Rechtlicher Rahmen	Auftrag 9 · Messartefakte erkennen	Auftrag 10 · Übertragbarkeit und Testdesign
Welche Quoten berichten Onlineshops in DACH/EU real — und unter welchen Voraussetzungen?	Wie viel Prozent der Anfragen entfallen auf Versand-, Lieferstatus- und Rückgabe-Intents?	Wo endet Automatisierung rechtlich — automatisierte Entscheidung, Widerruf, Transparenz?	Mit welchen Artefakten werden Containment-Zahlen aufgebläht — und wie prüft man sie gegen?	Wie prüfen wir die 60 % für unseren Assistenten belastbar — Stichprobe, Urteile, Abbruchkriterien?
Abgrenzung: Keine Meinungsartikel, keine Marktprognosen. Quellen geprüft: 16 Quellen / 12 Domains · Chatarmin (140 Shops, 2,6 Mio. Tickets: Median 16,3 %), Juskys/cleanercode, OBI (20 %)	Abgrenzung: Keine Performance-Zahlen, keine Hochrechnung. Quellen geprüft: 15 Quellen / 11 Domains · arXiv:2603.29888 (56,5 % After-Sales, offener Nenner), Gorgias Ecom Lab (10.990 Brands)	Abgrenzung: Keine Rechtsberatung; nur Rahmen und Prüfpunkte. Quellen geprüft: DSK-OH Rz. 12–13/22 · § 356a BGB (gesetze-im-internet.de) · EU AI Act Art. 50 (Amtsblatt-Volltext)	Abgrenzung: Keine Zahlenbewertung ohne Metrik-Definition. Quellen geprüft: Intercom-Pricing-Doku (assumed resolutions) · Zendesk-Reporting-Scope · Platforma 2026 (Close-on-silence 72 h, Reopen)	Abgrenzung: Keine Entscheidung für den PO; nur prüfbarer Plan. Quellen geprüft: Brown/Cai/DasGupta 2001 → n ≥ 400 (bei 65 % untere Grenze 60,2 %) · τ-bench (pass^k) · ECom-Bench 43,9 %

Sprint in Scrum

Sprint Goal

Ergebnis und Sprint Backlog

Belastbare externe Evidenz zur 60-%-Schwelle liegt vor: 29 dokumentierte Datenpunkte, Quellenart und Fundstelle je Quelle, Gegenbelege gesucht, Übertragbarkeit bewertet. Backlog: 6 parallele Rechercheaufträge + 1 Nachsteuerung (Übergabe/QA) + Zusammenführung.

Definition of Done

Jede Zahl mit Link, Herausgeber, Jahr, Stichprobe, Messdefinition, Zeitraum, Fundstelle · fehlende Angaben OFFEN · Anbieter- von unabhängigen Angaben getrennt · Gegenbelege benannt · keine erfundenen Daten · Ablage im Wikivault.

Story Map und erste Scheibe

Ablaufprototyp mit Link und Ergebnissen der beiden Tests

Story Map: Anfrage stellen → Antwort mit Fundstelle lesen → bestätigen oder übergeben. Test 1 (Versanddauer): Antwort „drei Werktage“ mit FAQ als Quelle — bestanden. Test 2 (beschädigte Ware): Übergabe mit Verlauf an Mitarbeitende — bestanden mit Einschränkung (Trigger-Recall nicht gemessen).

Auswertung (Analyze)

Geprüfte Belegtabelle

Link, Befund und Prüfvermerk je Quelle

Ablage: 011 Weiterbildung/Agile-Heros-Ausbildung/003 KI Manager Ausbildung/ — „Tag 5 - Recherche Versand- und Rueckgabefragen“ (.md/.pdf, 8 Seiten) und „…-Assistent.xlsx“ (6 Blätter).
Prüfvermerk: 29 Datenpunkte — 5 unabhängig/wissenschaftlich, 11 Anbieterangaben mit eigener Metrik, 6 Praxisberichte (Zeitraum teils offen), 3 Amts-/Standard-, 4 Rechts- und Behördenquellen. Als nicht verifizierbar verworfen: „35–45 % WISMO (Zendesk CX Trends)“, „40 % WISMO (Forrester)“, „80 % repetitive Tickets“.

60-%-Schwelle

Externe Evidenz und eigene Wirkung getrennt bewerten

Externe Evidenz: als Normalfall für alle Anfragen NICHT gestützt — 14 % vollständig im Self-Service gelöst (Gartner, n = 5.728), DACH-Median 16,3 % (140 Shops / 2,6 Mio. Tickets), OBI 20 %. Als Best-Case für engen Intent-Block mit Live-Datenanbindung GESTÜTZT (Top-Quartil 61,2 %; dokumentierte Sprünge 1,8 %→~60 % und 25 %→~75 % nach Systemanbindung). Anbieterangaben 76–82 % sind nicht unabhängig bestätigt.
Eigene Wirkung: ohne Pilotdaten NICHT ENTSCHEIDBAR — prüfbar über n ≥ 400 stratifizierte Fälle (untere Wilson-Grenze bei 65 % beobachtet: 60,2 %), Doppelcodierung mit κ, plus False Containment, Recontact 72 h, Trigger-Recall.

Iteration Review am Nachmittag

Entscheidung des Product Owners mit Grund

Vorschlag (Entscheidung liegt beim PO): Hypothese beibehalten, aber als Best-Case-Annahme kennzeichnen und an den Pilotversuch binden. Grund: für das Gesamtvolumen extern nicht gestützt; erreichbar nur mit Live-Datenanbindung und engem Intent-Zuschnitt. Ohne Systemanbindung realistische Erwartung 20–25 %.

Zweite Welle

Lücke, Folgeauftrag, Befund und aktualisierte Entscheidung

Lücke: „Übergabe und Qualitätsprüfung“ war zunächst nicht eigenständig abgedeckt. Folgeauftrag: Nachsteuerung an Rechercheauftrag 6. Befund: Übergabe ist die schwächste Stelle (10 % „immer reibungslos“, 59 % erklären erneut); keiner dieser Fehler senkt die Quote, mehrere erhöhen sie.
Aktualisiert: Quote nur zusammen mit Trigger-Precision/-Recall und Recontact Rate werten.

Retrospective – eine Änderung der Arbeitsweise

Änderung: Der Zugriffsweg wird vor dem Start je Subagent fest vorgegeben (Reader/curl + pdftotext statt Suchtreffer), weil das Extraktionswerkzeug funktionslos war. Außerdem: Eine Zahl gilt erst als „belegt“, wenn Metrik, Stichprobe und Fundstelle dokumentiert sind — nicht nach Sichtung des Abstracts.